D386–D390 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 28 November 2006*
doi:10.1093/nar/gkl932

MiST: a microbial signal transduction database

Luke E. Ulrich[1,3,*] and Igor B. Zhulin[1,2,3]

[1]Joint Institute for Computational Sciences and [2]Graduate School of Genome Science and Technology, The University of Tennessee–Oak Ridge National Laboratory, Oak Ridge, TN 37831-6173, USA and [3]Center for Bioinformatics and Computational Biology, School of Biology, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA

Received August 15, 2006; Revised October 12, 2006; Accepted October 13, 2006

ABSTRACT

Signal transduction pathways control most cellular activities in living cells ranging from regulation of gene expression to fine-tuning enzymatic activity and controlling motile behavior in response to extracellular and intracellular signals. Because of their extreme sequence variability and extensive domain shuffling, signal transduction proteins are difficult to identify, and their current annotation in most leading databases is often incomplete or erroneous. To overcome this problem, we have developed the microbial signal transduction (MiST) database (http://genomics.ornl.gov/mist), a comprehensive library of the signal transduction proteins from completely sequenced bacterial and archaeal genomes. By searching for domain profiles that implicate a particular protein as participating in signal transduction, we have systematically identified 69 270 two- and one-component proteins in 365 bacterial and archaeal genomes. We have designed a user-friendly website to access and browse the predicted signal transduction proteins within various organisms. Further capabilities include gene/protein sequence retrieval, visualized domain architectures, interactive chromosomal views for exploring gene neighborhood, advanced querying options and cross-species comparison. Newly available, complete genomes are loaded into the database each month. MiST is the only comprehensive and up-to-date electronic catalog of the signaling machinery in microbial genomes.

INTRODUCTION

Microbial signal transduction (MiST) links environmental stimuli to specific adaptive cellular responses. Consequently, signal transduction pathways are vital to an organism's survival and function. In microorganisms, they control the majority of cellular functions including chemotaxis, respiration, development, osmoregulation, transport, metabolism, virulence, host-recognition and antibiotic resistance. Such signal transduction pathways behave as information processing circuits that link input (stimulus—nutrients, temperature, redox, etc.) and output (regulatory response—gene expression, enzyme activity, flagellar motor switch, etc.) events (1,2).

The detection of a signal (input) and coupling this with an adaptive cellular response (output) is common to all signal transduction systems; however, microorganisms employ diverse mechanisms for linking these events. These range from single-domain transducers to several interacting proteins and multi-protein complexes. The most widely recognized signaling systems are the so-called two-component systems that utilize protein phosphorylation as the fundamental signaling mechanism (2,3). The prototypical two-component system consists of two proteins: a membrane-bound, sensor histidine kinase and a cytoplasmic, response regulator. The sensor kinase detects environmental signals via one or more, amino-terminal sensory domains. Subsequently, the sensor kinase undergoes a conformational shift that results in an ATP-dependent autophosphorylation of a conserved histidine residue within its carboxy-terminal, transmitter domain. The cognate response regulator then catalyzes the transfer of this phosphoryl group to a conserved aspartate residue within its amino-terminal, receiver domain. Phosphorylation of the receiver domain activates the response regulator's output domain(s) and effects a particular adaptive response—typically regulation of gene expression at the transcriptional level (1).

Despite the importance of two-component systems, most signal transduction events in prokaryotes are carried out by one-component systems that consist of a single protein molecule containing both input and output domains but lacking the phosphotransfer domains typical of two-component systems (4). Many one-component systems have been extensively studied including the LacI lactose operon repressor (5) and the catabolite activator, CAP, of *Escherichia coli* (6); the arginine catabolism regulator, RocR, of *Bacillus subtilis* (7); and the quorum-sensing regulator, TraR, of *Agrobacterium tumefaciens* (8). These proteins bind ligands via their input domain and regulate gene expression with their output domain. Most one-component systems typically consist of at least one input and one output domain, yet in some cases a single domain is capable of both detecting a ligand and producing a regulatory response. For example, both the

*To whom correspondence should be addressed. Tel: +1 865 974 7687; Fax: +865 576 4368; Email: ulrichle@ornl.gov

metalloregulators, CzrA of *Staphylococus aureus* (9) and CmtR of *Mycobacterium tuberculosis* (10), consist of a single domain, which directly binds DNA in response to metal ligands. One-component systems are more widely distributed among bacteria and archaea, and display a greater diversity of domains than two-component systems (4).

Signal transduction proteins are highly modular with diverse and mosaic domain architectures. The domains comprising these proteins may be categorized into four major types according to their function: input, transmitter, receiver and output (1). Relatively few input domains specific to signal transduction have been characterized and include the following: PAS (11,12), GAF (13), Cache (14), CHASE (15,16), CHASE2 through CHASE6 (17), NIT (18) and 4HB_MCP (19). These domains exhibit extreme sequence variability due to the broad range of signals they detect and are the least conserved of all the signaling modules. In contrast, transmitter and receiver domains display remarkable sequence conservation, which reflects the conserved phosphorylation reaction between these domain types that links sensor kinases and response regulators (20). The most commonly found output domains are DNA-binding helix–turn–helix (HTH) domains because the predominant adaptive response of two- and one-component systems is regulation of gene expression (4). Several novel output domains have been recently described in response regulators, which implicate these systems in other types of control, such as the regulation of enzyme activity. These include adenylate and diguanylate cyclases, c-di-GMP-phosphodiesterase, phosphohydrolase and other related domains (21–23). Output domains are more conserved than input domains yet still considerably divergent due to their moderate number of regulatory roles (e.g. gene-regulation, protein–protein interactions, etc.).

One of the first efforts to catalog signal transduction proteins was the SENTRA database (24). It contains information on classical two-component systems and a few other signaling systems that interact via phosphorylation or methylation reactions. Currently, SENTRA includes information on two-component signal transduction proteins for 43 genomes. Several other databases attempt to document the signal transduction machinery within various genomes; however, these projects are usually a part of a larger initiative and therefore limited in scope and/or accuracy. For example, the KEGG project (25) focuses on deriving higher-order information (e.g. pathways) from genomic data and has successfully mapped metabolic pathways across multiple genomes, yet the curators have recently begun mapping some regulatory pathways including signal transduction. KEGG transfers pathway information among genomes by first defining reference pathways based on scientific literature and then computationally extending these networks across genomic data via orthologous relationships established from sequence similarity searches and the positional correlation of genes. This approach is constrained to experimentally defined pathways and does not provide a complete record of the signal transduction repertoire. In addition, KEGG defines orthologous relationships using bi-directional best hits with respect to the entire protein sequence—this poorly ascertains related signal transduction proteins, given their highly modular nature and extreme sequence variability. Model organism databases such as the *E.coli* EcoCyc (26) and *B.subtilis* DBTBS (27) often contain information about signal transduction for their associated organisms, since these are primarily restricted to literature-based curation efforts and therefore limited in coverage (albeit with high accuracy). Finally, as one-component systems have only recently been recognized as a major part of signal transduction (4), most database resources when describing signal transduction contain only information on two-component systems.

We have developed a novel resource, the MiST database that contains a comprehensive compilation of the signal transduction proteins within bacterial genomes. As these proteins are modular in nature, our approach focuses on determining signal transduction proteins from a protein's domain composition, which we derive using the HMMER software (28). From the Pfam (29) and SMART (30) domain libraries, we have designated a set of profile hidden Markov models (HMM) that represent signaling domains (e.g. HATPase_c transmitter domain or DNA-binding output domains) and implicate a role in signal transduction. We classify a protein as belonging to signal transduction if it contains one or more of these specific signaling domains. In this manner, we systematically produce the repertoire of signal transduction proteins for microbial genomes (or any other collection of protein sequences). To our knowledge, signal transduction is the only current resource on signal transduction in prokaryotes that provides a thorough catalog of both two- and one-component systems within bacterial genomes.

HIGH-THROUGHPUT IDENTIFICATION OF SIGNAL TRANSDUCTION PROTEINS

The biological function of a signal transduction protein is determined by specific signaling domains that detect an environmental cue (input), mediate protein–protein communication (transmitter, receiver), or initiate a cellular response (output). Thus, domains that perform these roles serve as markers and facilitate the identification of signal transduction proteins given two conditions: (i) an adequate mechanism for detecting domains in an amino acid sequence and (ii) a comprehensive set of domains known or predicted to participate in signal transduction. Protein domains may be robustly represented by profile HMMs (28), which statistically model the primary structure of homologous domain sequences and enable their rapid identification from a protein sequence with tools such as the HMMER software package. Furthermore, domain profiles are more sensitive than pairwise sequence comparisons and typically contain manually-curated score thresholds from which the significance and membership of domain matches may be automatically evaluated. To meet the second requirement, we selected 133 signaling domains (Supplementary Table S1) from the Pfam (29) and SMART databases (30) based on the following: (i) known domain function (information from Pfam, SMART, InterPro (31), COG (32) resources and analysis of literature on signal transduction), and (ii) predicted domain function based on the association with other signaling domains. This approach has been recently described in detail (4).

MiST incorporates a straightforward and systematic methodology for identifying signal transduction proteins within completely sequenced bacterial genomes (Figure 1).

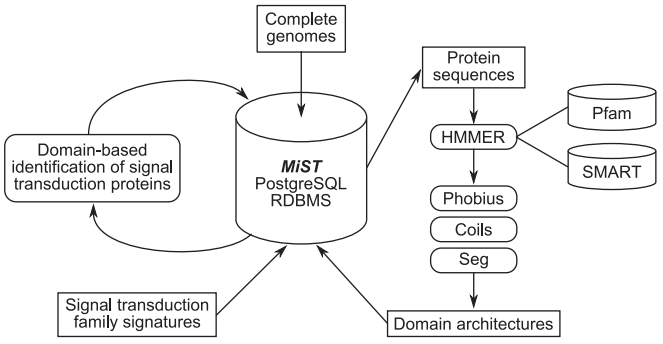

Figure 1. Overview of the high-throughput process for identifying signal transduction proteins. First, all complete, bacterial genomes are downloaded from NCBI and loaded into the MiST database. Second, the complete domain architecture of each protein is predicted. Finally, a protein is classified as belonging to signal transduction if it contains at least one transmitter, receiver, or output signaling domain.

Perl scripts download all available genomes from NCBI in the XML (eXtensible Markup Language) format and load this information into the database. Each XML file contains the full RefSeq annotation (33) for a genome, which includes its associated nucleotide data, genes and translated proteins. Secondly, we predict the domain architecture for each protein including signal peptides, transmembrane regions, coiled-coils and low-complexity segments. Finally, we identify the set of putative signal transduction proteins by scanning each protein for the presence of transmitter, receiver or output signaling domains (see above). Input domains are often found in pathways other than signal transduction (e.g. metabolic pathways) and therefore proteins identified solely from an input domain are not classified as belonging to signal transduction. During this process, we also filter out various domain combinations that indicate a role other than signal transduction. For example, due to structural similarities, DNA topoisomerase IV often contains a predicted HATPase_c transmitter domain (Pfam accession no. PF02518, SMART accession no. SM00387) in addition to other domains suggestive of topoisomerase activity (DNA_gyraseB and DNA_gyraseB_C), yet this protein is not involved in signal transduction. This pipeline may be executed on a regular schedule such that new genomes are seamlessly integrated into the database and signal transduction blueprints for these organisms are automatically generated. MiST is updated on a monthly basis to maintain a current record of the signal transduction repertoire within newly sequenced genomes.

The MiST database is implemented using the PostgreSQL (http://postgresql.org) version 8.1.3 relational database management system on the Gentoo distribution (http://www.gentoo.org) of the Linux/GNU operating system. All protein domain architectures are derived on our 34-node Linux cluster using HMMER version 2.3.2 (28) and the Pfam version 19.0 (29) and SMART version 5.0 (30) domain libraries. Every year, the domain architectures and signal transduction predictions will be updated based on the latest releases of the Pfam and SMART databases. Phobius version 1.01 (34) is used to predict signal peptides and transmembrane regions. Coiled-coils and regions of low-complexity are predicted with the COILS version 2.2 (35) and SEG (36) programs,

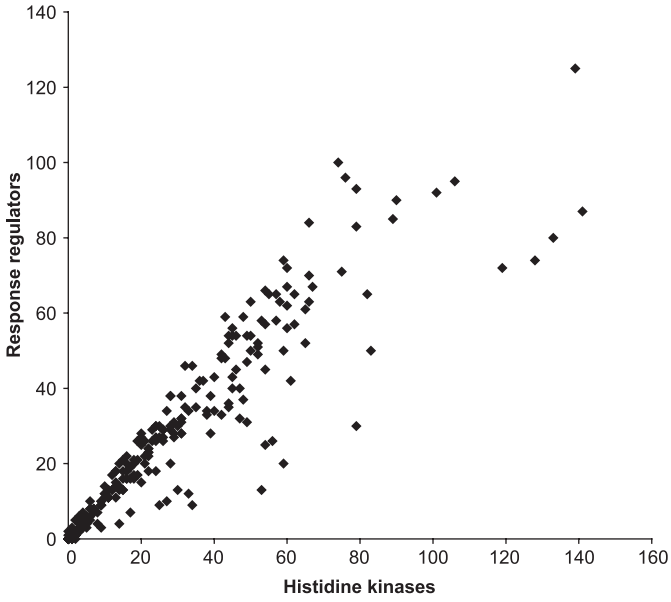

Figure 2. Scatterplot showing the linear, positive relationship ($R = 0.92$) between the number of predicted histidine kinases and response regulators within representative bacterial genomes (excluding chemotaxis proteins).

respectively. Custom Perl scripts handle program execution, formatting and interacting with the PostgreSQL subsystem. PHP and CGI scripts running on the Apache web server version 2.0.58 (http://apache.org) deliver web content to users via the Internet.

DATABASE CONTENTS

As of August 2006, MiST version 1.0 contains 69 720 predicted signal transduction proteins—22 868 proteins that belong to two-component regulatory systems and 46 402 one-component regulators—from 365 bacterial and archaeal genomes. Excluding chemotaxis proteins, there are almost equal numbers of sensor kinases and response regulators: 9317 (51.8%) and 9025 (49.2%), respectively. These display a strong, positive, linear relationship (Figure 2). Using several chemotaxis-related domains from the Pfam database, we found 4526 chemotaxis proteins, out of which 328 are putative Class II histidine kinases (CheA-like) (37). MiST contains more than twice as many one-component proteins as two-component proteins. When considering the number of systems based on their output activity, the number of one-component systems exceeds two-component systems by a factor of five. This is slightly higher than our initial analysis of the signaling transduction proteins derived from 145 prokaryotic genomes (4). The taxonomic distribution of two- and one-component systems within the MiST database is given in Supplmentary Table S2.

In addition to specific information about the signal transduction repertoire for each microbial genome, MiST stores the large amount of pre-computed data used to identify this class of proteins. This includes any predicted Pfam and SMART domains, protein secondary features (e.g. transmembrane regions, signal peptides, coiled-coils, low-complexity regions), chromosomal position of the corresponding gene,

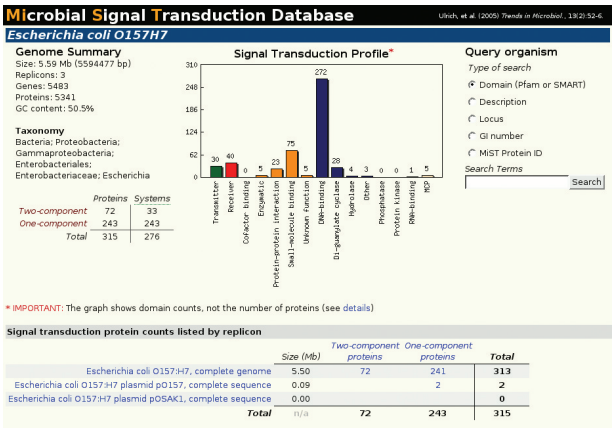

Figure 3. MiST screen-shot: the *E.coli 0157:H7* organism page. Clickable links are shown in blue. Queries can be made with various search types and terms (upper right corner).

statistical information and the complete RefSeq annotation. We have also indexed the NCBI non-redundant database to provide cross-referencing support. Therefore, while MiST focuses primarily on the classification and analysis of signal transduction proteins, this information along with its web interface (see below) enables users to effectively use MiST as a general resource for exploring microbial genomes.

WEB INTERFACE

To provide an effective means for exploring and utilizing the MiST database, we designed and implemented a user-friendly web interface oriented toward experimental biologists (http://genomics.ornl.gov/mist). A separate page for each microbe displays both general and specific information about its signal transduction repertoire including: descriptive data about the genome and proteome (e.g. size, replicons, etc.), a graphical representation of its signal transduction profile, querying options and a summary table of the signal transduction protein counts for each replicon (Figure 3). The profile represents an overall view of the number of signaling domains separated into functional categories. For example, *E.coli O157:H7* contains 272 DNA-binding domains, clearly indicating a substantial degree of gene-regulation. Each of the headings and signal transduction counts within the summary table are linked to the particular slice of data it represents. Clicking on a replicon displays the signal transduction proteins associated with that replicon. Selecting the heading 'Two-component proteins' displays all the predicted two-component proteins for this organism. Clicking on a signal transduction count will show either the two- or one-component proteins of a particular replicon. The resulting list for each protein contains GenBank Identifier (GI) number, RefSeq description, set of input and/or output domains and visualized domain architecture. Following an individual protein link provides an expanded overview of that particular protein and its associated gene. In addition to basic sequence data, this page contains the full protein/gene RefSeq annotation, domain architectures and graphical view of the gene neighborhood. The domain architecture section presents a linear representation of the protein's domain architecture and other secondary features (e.g. transmembrane regions, low-complexity regions). The chromosome view reveals neighboring genes as arrows pointed toward their transcriptional direction. Each arrow is hyperlinked to its respective gene/protein page for evaluating neighboring genes.

Relevant data may be accessed in several ways: (i) choosing an organism from the complete list of bacteria in MiST, (ii) searching for a particular organism, or (iii) searching by a GenBank GI number. In addition, multiple species may be selected and simultaneously searched using a sophisticated, taxonomy-driven, selection tool. Users may query for particular Pfam or SMART domains, keyword, locus, GI number or internal MiST identifier.

CONCLUSIONS AND FUTURE WORK

Using sensitive profile HMM searches and utilizing the available knowledge on microbial signal transduction from the literature and on-line resources, we were able to identify more than 69 000 proteins that are predicted to comprise signal transduction networks. We have developed a comprehensive database, which captures, stores and makes this information available to the scientific community. We expect MiST to become a widely-used resource significantly enhancing our knowledge within this important area of research as well as improving current genome annotation and contributing to important medical applications, such as antimicrobial drug development. Future MiST development will focus on incorporating user- and literature-based curation, GO ontology driven annotations (38), addition of new domain models, mapping functional residues, pathway support and improvements to the web interface.

SUPPLEMENTARY DATA

Supplementary Data are available at *NAR* online.

ACKNOWLEDGEMENTS

We thank the BLAST/STIM research community for many valuable discussions and suggestions on the MiST scope and development. We are grateful to Jan Lukens and 'Friends of the Zhulin lab' for their contribution toward our Linux cluster. This work was supported in part by a grant GM072285 from the National Institutes of Health to I.B.Z. and funds provided by the UT-ORNL Joint Institute for Computational Sciences. Funding to pay the Open Access publication charges for this article was provided by NIH grant GM072285.

Conflict of interest statement. None declared.

REFERENCES

1. Parkinson,J.S. and Kofoid,E.C. (1992) Communication modules in bacterial signaling proteins. *Annu. Rev. Genet.*, **26**, 71–112.
2. Stock,A.M., Robinson,V.L. and Goudreau,P.N. (2000) Two-component signal transduction. *Annu. Rev. Biochem.*, **69**, 183–215.
3. Hoch,J.A. and Silhavy,T.A. (1995) *Two-component signal transduction*. ASM Press, Washington, DC, USA.

4. Ulrich,L.E., Koonin,E.V. and Zhulin,I.B. (2005) One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.*, **13**, 52–56.
5. Lewis,M., Chang,G., Horton,N.C., Kercher,M.A., Pace,H.C., Schumacher,M.A., Brennan,R.G. and Lu,P. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, **271**, 1247–1254.
6. Kolb,A., Busby,S., Buc,H., Garges,S. and Adhya,S. (1993) Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.*, **62**, 749–795.
7. Calogero,S., Gardan,R., Glaser,P., Schweizer,J., Rapoport,G. and Debarbouille,M. (1994) RocR, a novel regulatory protein controlling arginine utilization in *Bacillus subtilis*, belongs to the NtrC/NifA family of transcriptional activators. *J. Bacteriol.*, **176**, 1234–1241.
8. Vannini,A., Volpari,C., Gargioli,C., Muraglia,E., Cortese,R., De Francesco,R., Neddermann,P. and Marco,S.D. (2002) The crystal structure of the quorum sensing protein TraR bound to its autoinducer and target DNA. *EMBO J.*, **21**, 4393–4401.
9. Pennella,M.A., Arunkumar,A.I. and Giedroc,D.P. (2006) Individual metal ligands play distinct functional roles in the zinc sensor *Staphylococcus aureus* CzrA. *J. Mol. Biol.*, **356**, 1124–1136.
10. Cavet,J.S., Graham,A.I., Meng,W. and Robinson,N.J. (2003) A cadmium-lead-sensing ArsR-SmtB repressor with novel sensory sites. Complementary metal discrimination by NmtR and CmtR in a common cytosol. *J. Biol. Chem.*, **278**, 44560–44566.
11. Ponting,C.P. and Aravind,L. (1997) PAS: a multifunctional domain family comes to light. *Curr. Biol.*, **7**, R674–R677.
12. Taylor,B.L. and Zhulin,I.B. (1999) PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol. Mol. Biol. Rev.*, **63**, 479–506.
13. Aravind,L. and Ponting,C.P. (1997) The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem. Sci.*, **22**, 458–459.
14. Anantharaman,V. and Aravind,L. (2000) Cache—a signaling domain common to animal Ca(2+)-channel subunits and a class of prokaryotic chemotaxis receptors. *Trends Biochem. Sci.*, **25**, 535–537.
15. Anantharaman,V. and Aravind,L. (2001) The CHASE domain: a predicted ligand-binding module in plant cytokinin receptors and other eukaryotic and bacterial receptors. *Trends Biochem. Sci.*, **26**, 579–582.
16. Mougel,C. and Zhulin,I.B. (2001) CHASE: an extracellular sensing domain common to transmembrane receptors from prokaryotes, lower eukaryotes and plants. *Trends Biochem. Sci.*, **26**, 582–584.
17. Zhulin,I.B., Nikolskaya,A.N. and Galperin,M.Y. (2003) Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in bacteria and archaea. *J. Bacteriol.*, **185**, 285–294.
18. Shu,C.J., Ulrich,L.E. and Zhulin,I.B. (2003) The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors. *Trends Biochem. Sci.*, **28**, 121–124.
19. Ulrich,L.E. and Zhulin,I.B. (2005) Four-helix bundle: a ubiquitous sensory module in prokaryotic signal transduction. *Bioinformatics*, **21** (Suppl. 3), iii45–iii48.
20. Grebe,T.W. and Stock,J.B. (1999) The histidine protein kinase superfamily. *Adv. Microb. Physiol.*, **41**, 139–227.
21. Galperin,M.Y., Nikolskaya,A.N. and Koonin,E.V. (2001) Novel domains of the prokaryotic two-component signal transduction systems. *FEMS Microbiol. Lett.*, **203**, 11–21.
22. Pei,J. and Grishin,N.V. (2001) GGDEF domain is homologous to adenylyl cyclase. *Proteins*, **42**, 210–216.
23. Shu,C.J. and Zhulin,I.B. (2002) ANTAR: an RNA-binding domain in transcription antitermination regulatory proteins. *Trends Biochem. Sci.*, **27**, 3–5.
24. Maltsev,N., Marland,E., Yu,G.X., Bhatnagar,S. and Lusk,R. (2002) SENTRA, a database of signal transduction proteins. *Nucleic Acids Res.*, **30**, 349–350.
25. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
26. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–D337.
27. Makita,Y., Nakao,M., Ogasawara,N. and Nakai,K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.*, **32**, D75–D77.
28. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
29. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
30. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
31. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
32. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
33. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
34. Kall,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
35. Lupas,A. (1996) Prediction and analysis of coiled-coil structures. *Methods Enzymol.*, **266**, 513–525.
36. Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
37. Bilwes,A.M., Alex,L.A., Crane,B.R. and Simon,M.I. (1999) Structure of CheA, a signal-transducing histidine kinase. *Cell*, **96**, 131–141.
38. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.